

Goodness-of-fit tests for correlated data

By THEO GASSER

Fachgruppe für Statistik, Eidgenössische Technische Hochschule, Zürich

SUMMARY

Goodness-of-fit tests for stationary processes are a problem of practical importance, e.g. in the analysis of electroencephalographic data. The distribution of the chi-squared statistic under the normal hypothesis is studied by simulation; power is investigated by an inverse filtering procedure for processes which can be well represented by an autoregressive-moving average model. For a second model, consisting of a Gaussian or non-Gaussian signal plus Gaussian noise, sample skewness and kurtosis are suggested as test statistics. The asymptotic normality and the asymptotic variance of these statistics are derived, as well as the behaviour for a broad class of alternatives. The second model is of primary interest in E.E.G.-analysis.

Some key words: Autoregressive-moving average process; Chi-squared test; Goodness-of-fit test; Kurtosis; Skewness; Stationary process; Time series.

1. INTRODUCTION

The classical goodness-of-fit tests depend on two assumptions:

- (i) the data $\{X_t: t = 1, \dots, n\}$ constitute a set of independent random variables;
- (ii) these random variables are identically distributed.

To study time series we shall keep assumption (ii), stationarity, but drop (i), restricting attention, however, to particular models.

The first such model consists of the class of processes with rational spectra, i.e. the mixed autoregressive-moving average processes of order (P, Q) :

$$X_t = \sum_{k=1}^P A_k X_{t-k} + \sum_{k=1}^Q B_k Z_{t-k} + Z_t, \quad (1)$$

where $\{Z_t\}$ is a process of independent and identically distributed random variables and the usual conditions for stationarity and invertibility are assumed.

A slightly more general model is the moving average representation of infinite order, the so-called general linear model. In practice this has to be truncated to a finite moving average. Model (1) is parsimonious compared to a finite moving average and to a pure autoregressive model. As a second and more general model, suppose that

$$X_t = S_t + Z_t, \quad (2)$$

where S_t is a Gaussian or non-Gaussian signal and Z_t is Gaussian noise. The main difference from (1) is that S_t , the interesting component of the process, may be concentrated in some frequency bands and have zero energy outside.

The motivation for this work comes from the analysis of electroencephalographic (E.E.G.) data, where we find a number of patterns with strong correlations. Physiological hypotheses led to the question whether the E.E.G. is Gaussian. As a first, yet incomplete, test for a stochastic process the amplitude distribution has been tested, usually by the χ^2 test (Saunders,

1963; Elul, 1969; Weiss, 1973). The question of the effect of the strong violation of independence on the χ^2 test has been ignored; because a violation of stationarity would invalidate the test for normality Elul (1969) analyzed short records of 2 sec. To obtain 'more' data, the continuous record was then digitized at the much too high frequency of 200 Hz, instead of 50–60 Hz at the most and this introduces additional correlation. The signals in E.E.G. analysis, as for example, the α rhythm, are often concentrated in quite narrow frequency bands. Model (2) seems to be more appropriate for this situation.

Sections 2 and 3 of the present paper study (1). A sampling experiment indicates the effects on the distribution under the null hypothesis and on the power of the chi-squared statistic. Modifications for correlated data are suggested and tested. Section 5 deals with the problem of testing for normality, particularly with (2).

2. DISTRIBUTION OF X^2 UNDER THE NULL HYPOTHESIS

If independent and identically distributed observations are grouped into K intervals with probabilities p_i^0 under the null-distribution, the test statistic

$$X^2 = \sum_{i=1}^K \frac{(n_i - np_i^0)^2}{np_i^0}$$

is asymptotically distributed as χ_{K-1}^2 . If p_i^0 depends on S parameters and if these are replaced by multinomial maximum likelihood estimators then X^2 is asymptotically χ_{K-S-1}^2 . This test is asymptotically distribution-free (Kendall & Stuart, 1967, p. 419–).

For 11 different parameter sets (A_k, B_k) , realizations of

$$X_t = \sum_{k=1}^P A_k X_{t-k} + \sum_{k=1}^Q B_k Z_{t-k} + Z_t$$

were generated with $t = 1, \dots, n+40$ and the X_t for $t = 41, \dots, n+40$ were used to eliminate boundary effects, with $n = 200$. The simplest distributions for this purpose are those that remain invariant under linear transformations: Z_t is chosen to be Gaussian to study the distribution under the hypothesis. The number of classes is $K = 10$, of equal probability size under the hypothesis, and the mean and variance are estimated by maximum likelihood estimators for nongrouped observations. All computations were performed at the Computer Center of the E.T.H. Zürich on a CDC 6400/6500 system.

As stated above, X^2 is asymptotically distributed as χ_{K-S-1}^2 if the parameters are determined as multinomial maximum likelihood estimators. Chernoff & Lehmann (1954) have shown that this is no longer true for maximum likelihood estimation from nongrouped data. The departures, however, decrease very quickly with increasing K . For $K = 10$ it may still explain to some extent the size of the tail probabilities for white noise, process 11, which are larger than expected.

The number of samples in the Monte Carlo study is $N = 1000$. Table 1 gives the number of rejections at the 10, 5 and 1 % significance level. As a measure for the overall fit to the χ_7^2 distribution ($K = 10$, $S = 2$), 10 groups of 100 samples are formed and a Kolmogorov–Smirnov test at the 5 % level is applied to each group.

Processes 1–6 are Markovian, both of low and of high frequency type, the spectrum of example 7 has a broad and weak peak at approximately 0.10 Hz and processes 8, 9 and 10 have a strong and narrow peak at approximately 0.35, 0.15 and 0.38 Hz respectively.

The following conclusions can be drawn from Table 1. For processes 1 and 7 with low to moderate correlation, the distribution of X^2 under the hypothesis does not change significantly. For strongly correlated processes, there is a sharp contrast between low and high frequency patterns; compare processes 3 and 4 with 5 and 6, and process 8 with 9. Series with high frequency correlation have small to moderate deviations, whereas low frequency patterns lead to gross deviations from a χ^2_7 distribution.

Table 1. *Distribution of X^2 under the normal hypothesis*

	Process	5 % K.S. rejects	Rejects at 10 %	Rejects at 5 %	Rejects at 1 %
	Ideal expected value	0.5	100	50	10
1	$P = 1, Q = 0: A_1 = 0.3$	2	95	46	10
2	$P = 1, Q = 0: A_1 = 0.6$	2	132	74	11
3	$P = 1, Q = 0: A_1 = 0.75$	7	171	98	21
4	$P = 1, Q = 0: A_1 = 0.9$	10	358	274	136
5	$P = 1, Q = 0: A_1 = -0.75$	2	108	57	7
6	$P = 1, Q = 0: A_1 = -0.9$	4	164	79	17
7	$P = 2, Q = 0: A_1 = 0.4, A_2 = -0.5$	3	109	48	11
8	$P = 2, Q = 0: A_1 = -1.1, A_2 = -0.85$	5	158	79	17
9	$P = 2, Q = 0: A_1 = +1.1, A_2 = -0.85$	2	125	64	17
10	$P = 2, Q = 2: A_1 = -1.2727, A_2 = -0.81, B_1 = 1.2727, B_2 = -0.81$	0	111	62	10
11	$P = 0, Q = 0$: white noise	2	116	62	10
	Thinned sequence				
3	$P = 1, Q = 0: A_1 = 0.75$	3	117	61	12
4	$P = 1, Q = 0: A_1 = 0.9$	6	117	64	5

Possible modifications to improve the approximation to χ^2_{K-S-1} for strongly correlated data include:

- (i) modification of the degrees of freedom (Patanakar, 1954);
- (ii) inverse filtering, see § 3;
- (iii) if proposal (ii) is not adequate, a simple but data-consuming procedure is thinning to reduce correlation. Associated with (iii) is the question of what sampling interval should be chosen to digitize the continuous record. The number of observations n_{mod} to be retained when thinning should have the following properties:

- (a) $n_{\text{mod}} \leq n$, with equality for white noise only;
- (b) when digitizing a continuous record of duration $T = n\Delta t$, n_{mod} should be asymptotically independent of the sampling frequency $1/\Delta t$ and proportional to T ;
- (c) the improvement should be such that standard techniques can be used, allowing for deviations from the assumed level of up to 20 % to fix ideas.

By not asking for optimality properties, one may run the risk of throwing away information for certain correlation patterns. The procedure to be proposed uses a simple functional of the spectrum, f , or the correlation function, ρ .

A. Estimate a number n_{mod} from

$$n_{\text{mod}} = \frac{\left\{ \int_{-\frac{1}{2}}^{\frac{1}{2}} f(\nu) d\nu \right\}^2}{\int_{-\frac{1}{2}}^{\frac{1}{2}} f^2(\nu) d\nu} = \frac{n}{\sum_{t=-\infty}^{\infty} \rho^2(t)}.$$

B. Round n_{mod}/n up to the next integer, which gives the spacing of the values to be retained.

This proposal satisfies our requirements. Table 1 shows that it brings the tail probabilities down to the values of the independent case for the critical examples $A_1 = 0.75$ and 0.9 , with n such that $n_{\text{mod}} = 200$.

The heuristic background lies in a statistical analogy. Individual periodogram values are asymptotically distributed as χ^2_2 . An average of n periodogram values of a nonwhite spectrum is not distributed as χ^2_{2n} , but approximately as χ^2 with $2n_{\text{mod}}$ degrees of freedom.

3. POWER

It is heuristically clear from a central limit argument that a goodness-of-fit test applied to a linear process

$$X_t = \sum_{k=-\infty}^{\infty} a_k Z_{t-k}$$

may lose much power as compared with direct observation of the white noise process Z_t . The distribution of X_t is closer to normal than the distribution of Z_t (Mallows, 1967). Given standardized Z_t with an absolutely continuous distribution G with finite third moment and $\{a_k\}$ standardized as

$$\sum_{k=-\infty}^{\infty} |a_k|^2 = 1,$$

then

$$\int_{-\infty}^{+\infty} \{F(y) - \Phi(y)\}^2 dy \leq g \max_k |a_k|^2,$$

where F is the distribution function of X_t , Φ is the standard Gaussian distribution function, and g is a constant, depending on G .

The hypothesis is that the X_t are normal. To judge the loss of power, a set of distributions G for Z_t with various characteristics is chosen: (i) G , uniform; (ii) G , double exponential; (iii) G distributed as χ^2_2 to represent asymmetry.

The Monte Carlo sample size is again $N = 1000$, with autoregressive-moving average process realizations of length $n = 200$ as in §2. Results are presented in Table 2. The estimated power of the independent case is given in the last row, power being the number of rejections $\times 10^{-3}$. Strong correlation leads to low power and the effect is more drastic for high frequency correlation patterns. The rows 'inversely filtered' are explained below.

If the distribution G of the input noise Z_t is long-tailed, double exponential for example, the loss of power is slightly less severe. Strong correlation patterns of high frequency type still lead to a reduction of power by up to a factor of 6. Details may be obtained from the author. For G asymmetric, the loss of power for low frequency correlation patterns is slight. For high frequency patterns it again becomes large. A remedy for this breakdown of power for model (1) is straightforward. It improves at the same time the approximation to the distribution of the test statistic under the hypothesis:

(i) the parameters of the autoregressive-moving average process are estimated by one of the established identification methods (Box & Jenkins, 1970, Chapter 6);

(ii) with the estimated parameter vectors \hat{A} and \hat{B} we inversely filter the series X_t to \hat{Z}_t so that it becomes approximately white noise, i.e. gives independent identically distributed values;

(iii) the χ^2 test is applied to \hat{Z}_t .

This was done for some of the cases in Table 2. For step (i) the following approximate maximum likelihood estimates for normal autoregressive processes of order 1 and 2 were used, in spite of the nonnormality of the Z_t :

$$\begin{aligned} \text{order 1: } \hat{A}_1 &= \hat{\rho}_1; & \text{order 2: } \hat{A}_1 &= \hat{\rho}_1(1 - \hat{\rho}_2)/(1 - \hat{\rho}_1^2), \\ & & \hat{A}_2 &= (\hat{\rho}_2 - \hat{\rho}_1^2)/(1 - \hat{\rho}_1^2), \end{aligned}$$

where $\hat{\rho}_i$ is the estimate of autocorrelation of lag i . The determination of the order of autoregression is a separate problem; the true order was assumed. The results are in Table 2 in the rows 'inversely filtered', for the process of the row above. There is a good improvement of power in all cases.

Table 2. Power of normal against uniform G_1 and against G_2 , namely χ^2_8

Process	Rejections at 10 % level		Rejections at 5 % level		Rejections at 1 % level	
	G_1	G_2	G_1	G_2	G_1	G_2
$P = 1, Q = 0: A_1 = 0.3$	655	914	497	848	244	688
$P = 1, Q = 0: A_1 = 0.6$	158	891	91	835	27	687
Inversely filtered	860	—	741	—	459	—
$P = 1, Q = 0: A_1 = 0.9$	292	1000	210	1000	93	1000
$P = 1, Q = 0: A_1 = -0.6$	—	369	—	258	—	108
Inversely filtered	—	958	—	934	—	845
$P = 1, Q = 0: A_1 = -0.9$	162	—	94	—	30	—
Inversely filtered	867	—	777	—	484	—
$P = 2, Q = 0: A_1 = 0.4,$ $A_2 = -0.5$	—	504	—	396	—	168
Inversely filtered	—	956	—	934	—	827
$P = 2, Q = 0: A_1 = 1.1,$ $A_2 = -0.85$	143	—	68	—	11	—
Inversely filtered	863	—	723	—	438	—
$P = 2, Q = 2: A_1 = -1.2727,$ $A_2 = -0.81, B_1 = 1.2727,$ $B_2 = -0.81$	91	302	40	218	12	128
$P = 0, Q = 0$: white noise	881	966	746	941	448	843

4. LIMITATIONS OF THE MODEL

Two types of deviations can jeopardize model (1) and as a consequence the inverse filtering procedure (i)–(iii). First, there may be a generation law of the type $Z_t \rightarrow X_t$ with Z_t an independent stationary process, but the transformation may be nonlinear. A general identification algorithm for nonlinear systems is not available. That any wide sense stationary process with absolute continuous integrated spectrum can be linearly transformed to an uncorrelated process is of no help and may obscure the real problem. In many cases where a nonlinearity exists, it is, however, of small order and we can neglect it to a first approximation.

Secondly, a serious objection is the following: X_t may not have a representation of the form $Z_t \rightarrow X_t$ with Z_t white noise, and with the transformation invertible. This is for example the case if the spectral density of X_t is zero on a set with positive measure. It is

obvious that inverse filtering is then not feasible. It is more dangerous when this situation holds for a component S_t of X_t , the 'signal process', on which is superposed coloured or white noise of a different kind. This occurs in E.E.G. analysis with Gaussian noise and a possibly non-Gaussian signal concentrated in some frequency bands. Inverse filtering would blow up the noise and pull down the signal, and a test in the Z_t domain would indicate normality even in cases of a non-Gaussian signal. This is the reason for considering model (2) in §5.

5. TEST FOR NORMALITY BY CUMULANTS

As outlined in §1, E.E.G. analysts interested in testing their data for normality have mainly used the χ^2 test, and in a few cases the Kolmogorov–Smirnov test. Neither is a particularly good choice for the specific hypothesis of normality (Shapiro, Wilk & Chen, 1968). There exist, however, powerful and at the same time simple tests for the particular case of the normal distribution. An attractive possibility is to use suitably normalized 3rd and 4th order cumulants. They are easy to evaluate, are location and scale invariant and have excellent power for a broad class of continuous alternatives. If we are confronted with the problem of correlated data, they have the additional advantage that they generalize quite easily. For a large sample $\{X_k: k = 1, \dots, n\}$ from a stationary process X_t , 3rd and 4th order sample cumulants are defined in the usual way, the difference between k statistics and sample cumulants being negligible. Then

$$\begin{aligned}\hat{k}_3 &= n^{-1}s_3 - 3n^{-2}s_2s_1 + 2n^{-3}s_1^3, \\ \hat{k}_4 &= n^{-1}s_4 - 4n^{-2}s_3s_1 - 3n^{-2}s_2^2 + 12n^{-3}s_2s_1^2 - 6n^{-4}s_1^4, \\ s_J &= \sum_{k=1}^n X_k^J.\end{aligned}$$

The application of the test statistic in a large-sample situation is based on the following theorem.

THEOREM. *Suppose that there is a Gaussian process $\{X_t\}$ ($t = 0, \pm 1, \dots$), $E(X_t) = 0$, with covariance function R_k and spectrum $F(\nu)$, with*

$$\sum_{k=0}^{\infty} |R_k| < \infty.$$

Then

(i) $n^{1/2}\hat{k}_3$ and $n^{1/2}\hat{k}_4$ are asymptotically jointly normal with expectation zero and finite variances and covariances;

$$\begin{aligned}\text{var}(\hat{k}_3) &= 6n^{-1} \sum_{k=-\infty}^{\infty} R_k^3 + O(n^{-2}), \\ \text{var}(\hat{k}_4) &= 24n^{-1} \sum_{k=-\infty}^{\infty} R_k^4 + O(n^{-2}), \\ \text{cov}(\hat{k}_3, \hat{k}_4) &= 0.\end{aligned}$$

Assertion (i) can be proved by appealing to a mixing condition. For this situation, however, a theorem by Sun (1965) is more specific. Some straightforward manipulations are needed to show the sufficiency of our condition. A complete proof of assertion (ii) would involve some tedious calculations, based on the decomposition of higher-order moments into moments of order two.

The following assumptions are needed to characterize the properties of \hat{k}_3 and \hat{k}_4 for a broad class of alternatives; l_m denotes the m th order cumulant.

ASSUMPTIONS. For $m = 2, 3, \dots$, (I) $\sum_{k_1} \dots \sum_{k_{m-1}} |c_m(k_1, \dots, k_{m-1})| < \infty$,

(II) $\sum_{k_1} \dots \sum_{k_{m-1}} |k_j c_m(k_1, \dots, k_{m-1})| < \infty$

with $j = 1, \dots, m-1$ and $c_m(k_1, \dots, k_{m-1}) = l_m(X_0, X_{k_1}, \dots, X_{k_{m-1}})$.

THEOREM. Assume that X_k is a strictly stationary process, with finite moments of arbitrary order. Then

(A) given assumption I, \hat{k}_3 and \hat{k}_4 are asymptotically unbiased and consistent estimates of $c_3(0, 0)$ and $c_4(0, 0, 0)$;

(B) given assumption II, \hat{k}_3 and \hat{k}_4 are asymptotically jointly normally distributed.

Proof. For simplicity, assume $E(X_k) = 0$. Using stationarity we have, for $F = E$ or l ,

$$\frac{1}{n^L} \sum_{i_1=1}^n \dots \sum_{i_L=1}^n F(X_{i_1}^{k_1} \dots X_{i_L}^{k_L}) = \frac{1}{n} \sum_{j_1=-(n-1)}^{n-1} \dots \sum_{j_{L-1}=-(n-1)}^{n-1} \left(1 - \frac{|j_1|}{n}\right) \dots \left(1 - \frac{|j_{L-1}|}{n}\right) F(X_{j_1}^{k_1} \dots X_{j_{L-1}}^{k_{L-1}} X_0^{k_L}).$$

This formula, together with the definition of cumulants in terms of moments and assumption I, leads to unbiasedness. For the proof of consistency we note that the assumption I also holds for $Y_i = (X_i - \bar{X})$. By the fact that those partitions that are not indecomposable cancel and by appealing to assumption I, we obtain consistency. The proof of (B) is most elegantly based on the asymptotic normality of linear estimates of 3rd and 4th order polyspectra (Brillinger & Rosenblatt, 1967).

Sample skewness and kurtosis are obtained by normalization:

$$\hat{\beta}_1 = \hat{k}_3 / \hat{\sigma}^3, \quad \hat{\beta}_2 = \hat{k}_4 / \hat{\sigma}^4.$$

COROLLARY. The quantities $n^{1/2} \hat{\beta}_1$ and $n^{1/2} \hat{\beta}_2$ are asymptotically jointly normally distributed with

$$\text{var}(\hat{\beta}_1) = 6n^{-1} \sum_{k=-\infty}^{\infty} R_k^3 + O(n^{-2}),$$

$$\text{var}(\hat{\beta}_2) = 24n^{-1} \sum_{k=-\infty}^{\infty} R_k^4 + O(n^{-2}),$$

$$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = O(n^{-2}).$$

The simple proof of this corollary is omitted.

6. THE APPLICATION OF $\hat{\beta}_1$ AND $\hat{\beta}_2$ TO STATIONARY PROCESSES

As in the independent case, the normal approximation can be safely used only for large sample size. A simple way to test the simultaneous hypothesis $\beta_1 = 0, \beta_2 = 0$ approximately at a level α is to do the tests individually at levels $\frac{1}{2}\alpha$. The first step consists in estimating $\Sigma \rho_k^3$ and $\Sigma \rho_k^4$, by truncating to a finite lag and inserting correlation estimates. The choice of the truncation point should be guided by a careful analysis of the empirical correlation

function and should be independent of n . The use of β_1 and β_2 is suggested as a preliminary test statistic for model (2), where inverse filtering is impossible when the signal process has its energy concentrated in some frequency bands. The application to a broad class of E.E.G. samples revealed a frequent violation of the hypothesis of normality of the amplitude distribution. Related to skewness and kurtosis are polyspectra which provide a test of normality not only of the first-order distribution of the stochastic process. The bispectrum for example is the spectral decomposition of the mixed third cumulant. For an application to E.E.G. data, see Dumermuth *et al.* (1971). Polyspectra have the additional advantage of being quite insensitive to violations of stationarity occurring in E.E.G. analysis. The need for large data size and relatively sophisticated mathematics may inhibit the widespread use of polyspectra.

REFERENCES

- BOX, G. E. P. & JENKINS, G. M. (1970). *Time Series Analysis, Forecasting and Control*. San Francisco: Holden Day.
- BRILLINGER, D. R. & ROSENBLATT, M. (1967). Asymptotic theory of estimates of k -th order spectra. In *Spectral Analysis of Time Series*, Ed. B. Harris, pp. 153–88. New York: Wiley.
- CHERNOFF, H. & LEHMANN, E. L. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Ann. Math. Statist.* **23**, 579–86.
- DUMERMUTH, G., HUBER, P. J., KLEINER, B. & GASSER, T. (1971). Analysis of the interrelations between frequency bands of the EEG by means of the bispectrum. *Electroenceph. Clin. Neurophysiol.* **31**, 137–48.
- ELUL, R. (1969). Gaussian behavior of the electroencephalogram during performance of mental task. *Science* **164**, 328–31.
- KENDALL, M. G. & STUART, A. (1967). *Advanced Theory of Statistics*, Vol. 2. London: Griffin.
- MALLOWS, C. L. (1967). Linear processes are nearly Gaussian. *J. Appl. Prob.* **4**, 313–29.
- PATANAKAR, V. N. (1954). The goodness of fit of frequency distributions obtained from stochastic processes. *Biometrika* **41**, 450–62.
- SAUNDERS, M. G. (1963). Amplitude probability density studies on alpha and alpha-like patterns. *Electroenceph. Clin. Neurophysiol.* **15**, 761–7.
- SHAPIRO, S. S., WILK, M. B. & CHEN, H. J. (1968). Comparative study of various tests for normality. *J. Am. Statist. Assoc.* **63**, 1343–72.
- SUN, T. C. (1965). Some further results on central limit theorems for non-linear functions of a normal process. *J. Math. & Mech.* **14**, 71–85.
- WEISS, M. S. (1973). Non-Gaussian properties of the E.E.G. during sleep. *Electroenceph. Clin. Neurophysiol.* **34**, 200–2.

[Received May 1974. Revised March 1975]